DOCUMENT RESUME

ED 121 280 IR 003 257

AUTHOR Geisinger, Kurt F.
TITLE A Systems Approach to Item Production and Review in a Computer Managed Instruction Project.
PUB DATE Apr 76
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, California, April 19-23, 1976)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage
DESCRIPTORS *Computer Assisted Instruction; *Criterion Referenced Tests; Higher Education; Inservice Education; *Item Banks; Nursing; Program Descriptions; *Test Construction; University Extension
IDENTIFIERS Computer Managed Instruction; Computer Managed Review and Examination; *Item Generation

ABSTRACT

An item generation procedure is described which was utilized in the development of Computer Managed Review and Examination courses for the education of nurses in remote areas. The major emphases are the processes of domain definition, item writing, and item edition. Specific discussion is presented concerning methods of item construction to assess technical vocabulary, concept learning, and the application of nursing principles to the solution of problems. The entire test construction procedure is briefly reviewed; this procedure includes numerous quality checks to insure the production of both high caliber instructional materials and domain-referenced tests. The criteria used at various editing and review stages are mentioned. An initial evaluation of the items is made, and problems inherent in the item generation procedure are offered. (Author/JY)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED121280

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A SYSTEMS APPROACH TO ITEM PRODUCTION AND REVIEW IN A COMPUTER MANAGED INSTRUCTION PROJECT*

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: IR TM
In our judgement, this document is also of interest to the clearinghouses noted to the right. Indexing should reflect their special points of view.

Kurt F. Geisinger
The Pennsylvania State University

Abstract

This paper describes the item generation procedure utilized in the development of Computer Managed Review and Examination courses for the education of nurses in remote areas. The major emphases are the processes of domain definition, item writing, and item editing. Specific discussion is presented concerning methods of item construction to assess technical vocabulary, concept learning, and the application of nursing principles to the solution of problems. The entire test construction procedure is viewed; this procedure includes numerous quality checks to insure the production of both high calibre instructional materials and domain-referenced tests. The criteria used at the various editing and review stages are mentioned. An initial evaluation of the items is made and problems inherent in the item generation procedure are offered.

Since the mid to late 1960's, traditional achievement testing has been the subject of considerable criticism and innovation. Glaser (1963), Bormuth (1970), and other measurement experts have strongly encouraged the educational community to re-evaluate the testing procedures used in instructional programs. The construction and selection of achievement test items, in particular, has been a focus of attention.

The problematic nature of achievement test and item construction rises to even greater prominence in Computer Managed Instruction (CMI) and other individualized instruction programs (Anderson et al.: 1974,

IR003257

*Presented at the 1976 AERA Annual Meeting, San Francisco, California, April, 1976.

Mitzel, 1974). As Robert Gagné (1975, p. 145) has said, "To the extent that modern educational trends, at all levels, favor self-education and independent learning, the means of observing and assessing the outcomes of learning becomes a matter of considerable importance."

The importance of item construction is exemplified in the fact that this CMRE project involves the development of eight nursing courses, each of which necessitates the construction of approximately 1,100 items[1], for a project total of almost 9,000 items!

Typically, testing in education serves a single prime purpose: to accredit or certify competence. Within the CMRE model of this project, however, four distinct purposes can be delineated for testing. First, the initial test serves a placement function. Then, the review questions throughout the instructional program serve two functions: to diagnose student learning and prescribe remedial instructional materials (and hence, to keep students from taking tests before they are adequately prepared for them) and to maintain student interest, motivation, and attentiveness. After instruction has been completed, the final examination serves the traditional credit awarding function.

Achievement measures are evaluated in terms of content validity. Most authors in the field of measurement (Cronbach, 1971) recognize that content validity is assessed largely with respect to the degree

[1]Of these 1,100 items, approximately 700 are test items and 400 are for review. Although the differences between these items is often not large, they are kept as two separate item pools. Since review items follow instruction more closely chronologically, we have tended to use these items with more specificity.

to which logical and systematic test construction procedures are utilized. The prime goal of this paper is to state the test construction procedures employed explicitly. Additionally, many individuals (Baker, 1974; Ferguson, 1972; Glaser, 1970; Hambleton, 1974; Hitke, 1974; Millingham and Geisinger, 1976) have been interested in the degree of parallelism between educational measurement practices and the other components of instructional systems. Only through such explicit statements of procedures can procedural inconsistencies ("working at cross-purposes") be identified and removed.

Developing the CMRE Courses

Staffing

Individual faculty members from the College of Human Development's Department of Nursing were assigned as course authors for each of the eight CMRE courses. Additionally, a nurse-research assistant was hired one-half time to assist each author. While these assistants aided the authors in reviewing academic materials for use "off-line," their major role was that of item writing. A professor from the Department of Educational Psychology conducted several item writing workshops for these individuals. Two graduate students in Educational Psychology served as principal item reviewers or editors, and wrote items upon occasion.

The Course Development Process

The following eleven step course development sequence was utilized in the preparation of all eight courses to facilitate progress and to

keep track of the location of items within the organization.

1. [Instructional Material Developed]
2. [Initial Test Item Construction]
3. [Initial Item Review]
4. First Item Revisions Made
5. Item Typing on Paper
6. Second Item Review
7. Second Item Revisions Made
8. Magnetic Tape Selectric Typing
9. Course Author Item Approval Granted
10. Magnetic Tape Corrected
11. On-line Review and Revision

Instructional Material Developed

The development of instructional materials for these courses is primarily a two-step process. Each author first enumerates his major goals for the course. From this list of goals, a detailed subject-matter outline is constructed. Using this outline, a committee of the Department of Nursing then judges the adequacy of course coverage. The second step involves the operationalization of the outline; appropriate materials [texts, articles, films, tapes, etc.] are developed or selected to represent the topic areas listed in the outline. [This procedure is crucial for the test construction process; Appendix B describes how the instructional material selection and development relates to current measurement topics such as universe and domain definition, domain-referenced testing, and criterion-referenced testing.] Then, this body of curricular information is divided into single class session-sized segments called lessons. Each lesson is weighted according to its importance, this weighted importance being directly proportional to the eventual number of items for that lesson.

The Use of Summary Statements

Consistent with the procedures used at the University of Illinois'

[illegible] project, subject matter experts derive summary statements from the instructional material [illegible] [illegible, 1972]. Summary statements are descriptions of the major ideas which instruction and of these elements of critical importance within such lessons. A single lesson might have as many as twenty to twenty-five such statements. These summary statements are written sentences from the textual material or which the lesson is based. The specific subject and predicate are exchanged from the exact wording of the text, the summary statements are cast in consistent with the language of the text as possible. The length of the summary statements ranges from a single sentence to a short paragraph.[1] Each summary statement is to be analyzed by subjects, lessons, and page of text in article-within the lesson.

Writing the Test Items

After preparing the instructor objectives representing activities, the subject matter experts involved in each course construct the bulk of the questions for that particular course. The subject matter experts, both course authors and course assistants are [illegible] Each item is constructed from a summary statement. [illegible] each summary statement, each item is referenced to the page of the article to [illegible] from which to work. This is used later in the [illegible] diagnostic-prescriptive statements for the questions.

---

[1] In the courses of increasing difficulty, after the first year of the project, some course authors have dispensed with the use of summary statements. In most cases, the authors thought it unnecessary. Use these statements in the instructional materials themselves. This, of course, eases the task of copying the statement sentence from the book. Furthermore, it permits the author to select those topics or areas to develop items, while assigning the course assistant to actually construct the items from the selected passages. This, too, represents a time-saving.

Basically, two distinct types of rules (parts of the code) are in the tools. The first is used if the summary statement includes a specific term or name which the subject matter expert believes the student should be able to recall (rather than recognize). A constructed response item, in which the student is required to type in the term, is used. In this case, however, the code of the item is stored from the computer summary statement through the use of paraphrasing. Therefore, grammatical transformations are also performed on the paraphrased summary statement. This means that the item is semantically encoded in memory, not merely syntagmatically or phonologically encoded (see Anderson, 1976, for an in depth explanation of this point.) Item writers must be reminded often that this type of item is only appropriate for specific words or phrases. It is a relatively easy trick of item construction whereby items is to be used for instructions, it is simply not justified for testing these general concepts with deep principle. Basically, an item that has an infinite synonym are trivial to be taught or correct. Approximately ten per cent of our items are of this variety.

A second type of rule examines the ability to identify learning principles. Basically, each principle statement a number of points to the learning student which is appropriate under certain circumstances. The code of such a question represents one such specific circumstance. The scorer is asked what to do. The rules must correctly apply the principle to all the situations. Unless, as is infrequently the case, the desired response is embedded as a specific term, a definite decision format is used, with a number of possible answers listed as options.

Content Validity

As mentioned earlier, content validity is largely assessed in viewing the test construction process systematically, and judging how adequately the test items represent the domain. The domain has been carefully defined and summary statements have been made extremely consistent with the instructional materials. Then, the statements are paraphrased and often grammatically altered such that a student must comprehend the instructional material to answer it correctly. Quality control is assured in that all items are read several times by several different people before the items go "on-line." A faculty committee of the Department has evaluated all course outlines as adequate representations of the subject matter. Furthermore, nationally known nursing experts are being brought to Penn State to evaluate the CMRE courses.

Problems

The test construction process appears to be largely successful. However, several problems do appear worthy of mention.

1. The fact that the item editors were largely unable to make nursing-related statements leads to inefficiency: this is especially troublesome in the attempt to generate plausible alternatives for the mulitple choice items.
2. Even with the utilization of carefully chosen off-line instructional material, academic idiosyncracies on the part of the text authors are found. In constructing questions to test

such material, the item writer is forced to preface the item with, "According to . . ." One of the goals of CMRE is to allow students who have previously learned material to bypass the instruction on it a second time. Material which is textbook - or author-specific is not conducive for this purpose.

3. When starting with summary statements, a writer finds that he can generate a considerable number of items from a single summary statement. Selecting which item is best is an extremely unscientific process. This is especially difficult when the different items appear to have widely different levels of difficulty.
4. The procedure of determining what an item assesses (a concept, a principle, a term, etc.) is a highly subjective, mentalistic process, subject to disagreement among item writers.

These are problems requiring practical solutions. As increasing numbers of CMI and CMRE projects are developed, we hope that such assessment problems can be solved or handled in a better manner. Of course, the ultimate beneficiaries of such solutions are not the future CMI developers, but the future students (Popham, 1974).

REFERENCES

1. Alkin, M. C. "Criterion-referenced measurement" and other such toms. In Harris, C. W., Alkin, M. C., and Popham, W. J. (Editors), Problems in criterion-referenced measurement, Center for the Study of Evaluation, University of California, Los Angeles, 1974.

2. Anderson, R. C. How to construct achievement tests to assess comprehension. Review of educational research, 1972, 42, 145-170.

3. Anderson, T. H., Anderson, R. C., Dalgaard, R. B. Wretecha, E. J., Biddle, W. B., Pader, D. W., Smock, H. R., Alessi, S. M., Surber, J. R., and Klemt, L. L. A computer based study management system. Educational Psychologist, 1974, 11, 36-45.

4. Baker, R. L. Measurement considerations in instructional product development. In Harris, C. W., Alkin, M. C., and Popham, W. J., (Editors), Problems in criterion-referenced measurement, Center for the Study of Evaluation, University of California, Los Angeles, 1974.

5. Bormuth, J. P. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

6. Cronbach, L. J. Test validation. In Thorndike, R. L. (Ed.), Educational measurement, (2nd ed.), Washington, D. C.: American Council on Education, 1971.

7. Davis, F. B. and Diamond, J. J. The preparation of criterion-referenced tests. In Harris, C. W., Alkin, M. C., and Popham, W. J., (Editors), Problems in criterion-referenced measurement. Center for the Study of Evaluation, University of California, Los Angeles, 1974.

8. Donlon, T. F. Some needs for clearer terminology in criterion-referenced testing. Presented at the annual meeting of the National Council for Measurement in Education, 1974.

9. Ebel, R. Some limitations of criterion-referenced measurement. Paper presented at the annual meeting of the American Educational Research Association, 1970.

10. Ferguson, R. L. A model for computer-assisted criterion-referenced measurement. Education, 1970, 81, 25-31.

11. Fremer, J. Development of school-based criterion-referenced testing systems. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D. C., April 2, 1975.

12. Gagne', R. M. Observing the effects of learning. Educational Psychologist, 1975, 11, 144-157.

13. Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

14. Glaser, R. Evaluation of instruction and changing educational models in Wittrock, M. C. and Wiley, D. C. (Editors). The evaluation of instruction: Issues and answers, New York: Holt, Rinehart, and Winston, Inc., 1970.

15. Glaser, R. and Nitko, A. J. Measurement in learning and instruction. In Thorndike, R. L., Educational Measurement, (2nd edition). Washington, D. C.: American Council on Education, 1971.

16. Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. Review of educational research, 1974, 44, 371-400.

17. Hambleton, R. K., Swaminathan, H., Algina, J. and Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. An invited symposium presented at the annual meeting of the American Educational Research Association, Washington, D. C., April, 1975.

18. Hively, W. Introduction to domain-referenced testing. Educational Technology, 1974, 14, 5-10.

19. Jackson, R. Developing criterion-referenced tests. TM Report No. 1, Princeton, New Jersey: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1970.

20. Klein, S. P. and Kosecoff, J. Issues and procedures in the development of criterion-referenced tests. TM Report No. 26, ERIC Clearninghouse on Tests, Measurement, and Evaluation, Educational Testing Service, Princeton, New Jersey, 1973.

21. Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

22. Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.) Evaluation in education: Current applications. Berkeley, California: McCutchen Publishing Co., 1974.

23. Mitzel, H. E. (Ed.) An examination at the short-range potential of computer-managed instruction. Conference Proceedings, November, 6-8, 1974.

24. Nitko, A. J. Problems in the development of criterion-referenced tests: The IPI Pittsburgh experience. In Harris, C. W., Alkin, M. C., and Popham, W. J. (Editors), Problems in criterion-referenced measurement. Center for the Study of Evaluation, University of California, Los Angeles, 1974.

25. Popham, W. J. Selecting objectives and generating test item for objectives-based tests. In Harris, C. W., Alkin, M. C., and Popham, W. I. (Editors), Problems in criterion-referenced measurement. Center for the Study of Evaluation, University of California, Los Angeles, 1974.

26. Popham, W. J. and Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

27. Shoemaker, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-147.

28. Skager, R. W. Generating criterion-referenced tests from objectives-based assessment systems: unsolved problems in test development assembly, and interpretation. In Harris, C. W., Alkin, M. C., and Popham (Editors), Problems in criterion-referenced measurement. Center for the Study of Evaluation, University of California, Los Angeles, 1974.

29. Tinkelman, S. N. Planning the objective test. In Thorndike, R. L., Educational Measurement, (2nd ed.) Washington, D. C.: American Council on Education, 1971.

30. Wesman, A. G. Writing the test item. In Thorndike, R. L., Educational Measurement, (2nd ed.) Washington, D. C.: American Council on Education, 1971.

31. Wietacha, E. J. and Anderson, R. C. The preparation of test items to maintain attentive study behavior. In Anderson, R. C. and Anderson, T. H., Development and Implementation of the CAI Study Management System (CAISMS). Preliminary technical report, March, 1975.

32. Willingham, W. W. and Getzinger, S. F. Developing an operational model #1: assessing experiential learning. In Willingham, W. W. and Knobitt, R. S., Implementing a program for assessing experiential learning. Princeton, New Jersey: Cooperative Assessment of Experiential Learning, Educational Testing Service, 1976.

33. Wood, D. A. Test construction. Columbus, Ohio: Merrill, 1961.

[redacted]

Generally, considerable inconsistency is observed between researchers' opinions regarding the use of criterion-referenced testing (Ebel, 1971; Gorham, 1971; Millman, 1973; Cox, 1974). Some researchers support the use of criterion-referenced testing, mainly since only in these tests does there exist an agreement between instructional objectives and test items. In the other hand, a criterion-referenced test is "any test exclusively or primarily useful for assessing an individual's performance with respect to a well-defined set of tasks" (Millman, 1974). In such a manner, each individual is measured in relation to the degree to which he has attained the mastery of instructional aims and not in relation to the performances of other examinees who happen to be currently in front of the instructional material. Millman further argues that such criterion-referenced tests yield scores which are unbiased estimates of the percentage of all items within a domain measured, or true scores. Such scores are commonly derived from well prepared and carefully developed, and for reporting student mastery. With a well-defined domain, there should be high agreement among experts as to what constitutes adequate items for the domain. Hambleton (1974) has argued that a "behavioral" or an "objective" should be specified sufficiently well enough to permit experts to match test items with the objectives. In this way, the specificity of the test items can be ascertained, as the set of "well-cast" instructional objectives would provide cues for test [illegible]

[illegible] criterion-referenced tests in instructional programs, see Cox & Vargas (1966). It is recommended that the reader refer to Popham & Husek (1969) for additional discussion of criterion-referenced measures and their use.